Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: January 8, 1998

Period of Report: October 1, 1997 to December 31, 1997

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

Distribution Statement A: Approved for public release; distribution is unlimited.

UNIVERSITY OF MASSACHUSETTS
AMHERST

Computer Science

Lederle Graduate Research Center
Box 34610
Amherst, MA 01003-4610
(413) 545-2744

DATE: January 8, 1998

TO: Defense Technical Information Center (DTIC)

FROM: W. Bruce Croft, Principal Investigator

SUBJECT: Quarterly Scientific and Technical Report for F19628-95-C-0235

Enclosed is your required number of copies of the quarterly R&D Status Report and Scientific and Technical Report for ARPA order number number D570 (note: changed from old AO #D468) issued by ESC/ENS under contract number F19628-95-C-0235. The title of the project is "Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents." These reports are being distributed in the appropriate amounts to ESC/AXS, ESC/ENK, ARPA/ITO, DTIC, and ARPA/Technical Library.

I have also enclosed a copy of the slides from the December meeting.

If you have any questions, I can be reached by email at croft@cs.umass.edu.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	01/08/98	Scientific/Tech

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D468

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR5281810198

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS			15. NUMBER OF PAGES
Browsing	Query Processing	Indexing	8
Image Retrieval	Scanned Document Retrieval	Bayesian Network	16. PRICE CODE
Text Retrieval	Probabilistic Retrieval Model	Large Distributed Databases	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we will be studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. Extensive use will be made of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query. We will also be making increased use of PTO text databases in these experiments.

Technical Results

The new phrase indexing approach was applied to patent data, This demonstrated that patents make heavy use of phrases and that the phrases are substantially different than those found in the TREC database (examples in overheads attached). A new patent retrieval demonstration incorporating this indexing was shown at the second DARPA/PTO status review meeting in Washington D.C. on the 10th December. In addition, for this demonstration another year of patents was indexed.

Important Findings and Conclusions

Initial results show that phrase indexing and query formulation techniques substantially improve the results of patent searches.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We plan to continue to enhance the query processing and retrieval strategies for patents, including the use of automatic query expansion techniques. We also plan a version of the patent search demonstration with an improved user interface to integrate Boolean and free text searching.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with

substantial input from PTO staff. Most of the classification experiments will be done in the context of the PTO classification and previously classified patents.

Technical Results

The TREC evaluation of an approach to visualizing retrieval results showed that some users were able to obtain significant retrieval benefits. A discussion took place at the December DARPA/PTO review about which visualizations may be the most useful for patent searching.

An on-line demonstration of the patent classification system was given at the December review meeting. This demonstration showed that nearest neighbor classification based on full patent searching can produce very good results.

Important Findings and Conclusions

The TREC evaluation was one of the first of this scale for this type of information visualization. Our results continue to indicate that many classes of patents could be reliably classified automatically.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We are now focusing on evaluating the classification accuracy and incorporating additional classification techniques into the classification system.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is "find me things that look

like this". We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web.

Technical Results

The new image retrieval techniques were applied to a much larger (60,000 image) trademark database and demonstrated in a multimodal text plus image trademark retrieval system at the December meeting. Some problems were found in that initial demonstration have since been fixed. Initial studies of flower patent retrieval were also presented at the meeting.

Important Findings and Conclusions

The feedback from the December meeting clearly indicated the benefits of combining text plus image retrieval. Important directions for improving the system were also discussed.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We will continue to improve the demonstration trademark retrieval system by refining the text search component and refining the image match algorithms. We also plan to further scale up the system to handle hundreds of thousands of trademarks.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient

retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

More experiments on distributed search were carried out. The DS3 connection to AAINet was finally installed in December.

Important Findings and Conclusions

None.

Significant Hardware Development

Special Comments

None.

Implications for Further Research

We are currently working on defining a distributed search experiment that would involve having one or more servers at the San Diego site.

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

December 1997

Status Report

W.B. Croft, L. Larkey, R. Manmatha

Center for Intelligent Information Retrieval

University of Massachusetts, Amherst

Overview of Tasks

- Representation techniques for complex documents
- Browsing and classification techniques for document collections
- Image indexing and retrieval
- Distributed retrieval architectures

Search Scenarios (Text and Image)

- Patent query -> Patent database
- Patent query -> External databases
- General query -> Patent database
- General query -> External databases

Representation Techniques

- Goal: Extend word-based representations to more effectively support summarization, browsing and retrieval
  - Subgoal: Exploit structure of patent documents
- Technical focus: Identifying phrases and phrase contexts, extending underlying retrieval model, query processing
  - Subfocus: Develop testbed using patent documents

Representation Techniques

- Developed lexical acquisition program for building a phrase dictionary from large databases
  - statistical approach faster and more accurate than part-of-speech tagging
  - heuristics needed to exclude uninteresting collocations
- Developed new class of operators for Bayesian Net model
  - enable more interesting combination of evidence than a linear weighted average
  - shown to be useful in modeling Boolean combinations

Phrase Extraction

- 1,100, 000 phrases extracted from all TREC data
  - more than 1,000,000 WSJ, AP, SJMS, FT, Ziff, CNN documents
- 3,700,000 phrases extracted from PTO 1996 data
- Currently used in query processing for patent retrieval demonstration

Top Phrases from TIPSTER

65824 United States
61327 Article Type
33864 Los Angeles
18062 Hong Kong
17788 North Korea
17308 New York
15513 San Diego
15009 Orange County
12869 prime minister
12799 first time
12067 Soviet Union
10811 Russian Federation
9912 United Nations
8127 Southern California
7640 South Korea
7620 end recording
7524 European Union
7436 South Africa
7362 San Francisco
7086 news conference
6792 City Council
6348 Middle East
6157 peace process
5955 human rights
5837 White House
5778 long time
5776 Armed Forces
5636 Santa Ana
5619 Foreign Ministry
5527 Bosnia-Herzegovina
5458 words indistinct
5452 international community
5443 vice president
5247 Security Council
5098 North Korean
5023 Long Beach
4981 Central Committee
4872 economic development
4808 President Bush
4652 press conference
4602 first half
4565 second half
4495 nuclear weapons
4448 UN Security Council
4426 South Korean
4219 first quarter
4166 Los Angeles County
4107 State Duma
4085 State Council
3969 market economy
3941 World War II

Top Phrases from Patents

975362 present invention
191625 U.S. Pat
147352 preferred embodiment
95097 carbon atoms
87903 group consisting
81809 room temperature
78458 SEQ ID
75850 BRIEF DESCRIPTION
66407 prior art
59828 perspective view
58724 first embodiment
56715 reaction mixture
54619 DETAILED DESCRIPTION
54117 ethyl acetate
52195 Example 1
52003 block diagram
46299 second embodiment
41694 accompanying drawings
40554 output signal
37911 first end
35827 second end
34881 appended claims
33947 distal end
32338 cross-sectional view
30193 outer surface
29635 upper surface
29535 preferred embodiments
29252 present invention provides
29025 sectional view
28961 longitudinal axis
27703 title compound
27434 PREFERRED EMBODIMENTS
27184 side view
25903 inner surface
25802 Table 1
25047 lower end
25047 plan view
24513 third embodiment
24432 control signal
24296 upper end
24275 methylene chloride
24117 reduced pressure
23831 aqueous solution
23618 SEQUENCE DESCRIPTION
23616 SEQUENCE CHARACTERISTICS
22382 weight percent
22070 closed position
21356 light source
21329 image data
21026 flow chart
21003 PREFERRED EMBODIMENT

Phrases from TREC Queries

14	international criminal activity
9	international criminal
1436	criminal activity
84	hubble telescope
188	passenger vehicle
9086	civil war
255	hydroelectric project
5261	detailed description
183	rap music
1449	negative effect
8081	young people
297	radio wave
26	radio tower
404	car phone
135	brain cancer
5	theft of trade secret
1324	trade secret
573	sources of information
530	trade journal
334	business meet
506	patent office
1870	trade show
26	competitor's product
63	growing plant
41	magnetic levitate
38	commercial harvest
58	highway accident

Representation Techniques

- Refined context-based query expansion
  - tested in recent TREC
- Initial evaluation of identifying "core concepts" in a query
  - also tested in TREC, being combined with new model
- Downloaded PTO Greenbook data and built database using INQUERY
  - includes all Greenbook fields, relevance feedback, query processing, various display enhancements

Clusters from Breast Cancer query

Group 1:
- breast cancer patient
- breast exam
- breast tissue
- u.s. women
- cancer kills
- cancer society
- cancer specialist
- family history
- mammogram
- mammography

Group 2:
- chemotherapy
- lumpectomy
- lymph node
- mastectomy
- radiation therapy
- recurrence
- survival rate

Group 3:
- breast implant
- implant
- silicone gel
- silicone gel breast implant
- silicone implant

Group 4:
- birth control pill
- breast cancer risk
- menopause
- sex hormone

Group 5:
- breast cancer surgery
- cancer surgery

Group 6:
- national cancer institute
- sloan kettering cancer center

Group 7:
- breast cancer research
- self examination

TREC Query Clusters

- For many queries, topic clusters are less clear
  - use alternate sources of topic hierarchies, e.g. Wordnet?
- Example TREC query about harmful effects of herbal food supplements

substance

disease

consumer food_product

nutrition_labeling drug content_claim drug_administration nutrition health_claim
nutrients herb mineral fda supplement vitamin food

listeria

herbs

Determining Core Concepts

- "What research is ongoing to reduce the effects of osteoporosis in existing patients as well as prevent the disease occurring in those unafflicted at this time?"
  - core concept: "osteoporosis"
- "Annual budget and/or cost involved with the management and upkeep of National Parks in the U.S."
  - "National Parks"
- Use combination of linguistic analysis, weighting, and corpus analysis of query word relationships

TREC Queries

```
#q307 = #WSUM ( 1
 1.0 #WSUM ( 1.0
1 project
1 construct
1 extent
1 desire
1 country
1 consequence
1 purpose
1 nature
1 hydroelectric
1.5 #foreigncountry
1 locate
1 propose
1.5 #passage25(  #PHRASE( hydroelectric project) )
)
```

```
 1.25 #WSUM(1.0
1 project
0.987143 construct
0.974286 dam
0.961429 #3( federal power act )
0.948571 #3( power project )
0.935714 #3( feasible study )
0.922857 ferc
0.91 #3( dam project )
0.897143 turbine
0.884286 #3( water manage )
0.871429 #3( rio arriba county )
0.858571 #3( mr. sharp )
0.845714 electric
0.832857 #3( construct license )
0.82 #3( ferc project )
0.807143 doe
0.794286 reclamation
0.781429 wcua
0.768571 #3( federal energy regulatory commission )
0.755714 commence
0.742857 laos
0.73 hungary
0.717143 #3( vinh son )
```

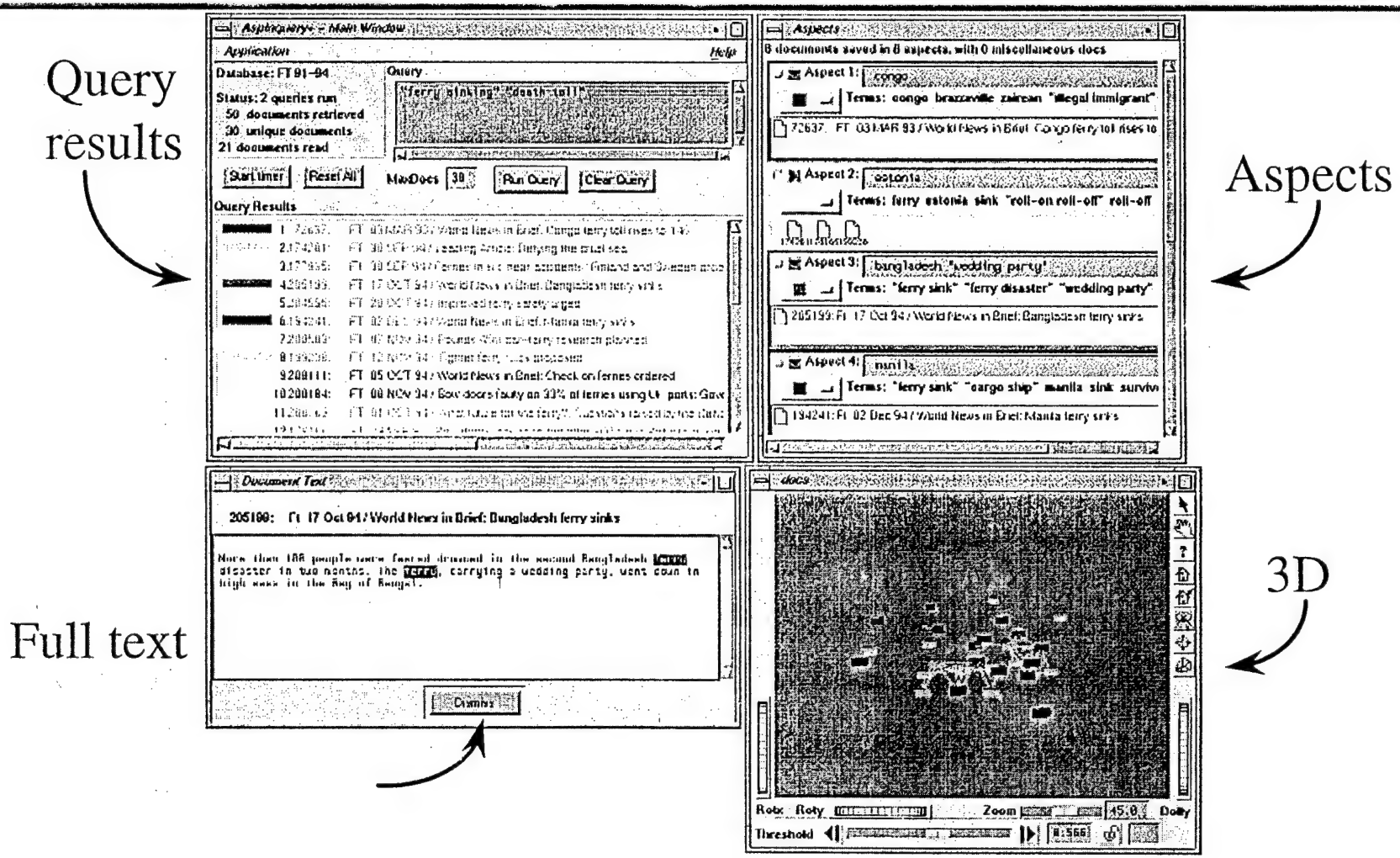
Browsing and Classification

- Goal: Develop techniques for classifying documents in order to improve effectiveness of interactive browsing and classification
  - Subgoal: Improve PTO classification structure and accuracy
- Technical Focus: Using clustering and 3-D visualization to summarize groups of documents; Using combinations of classification techniques to assign categories
  - Subfocus: Evaluate using TREC and PTO classification testbed

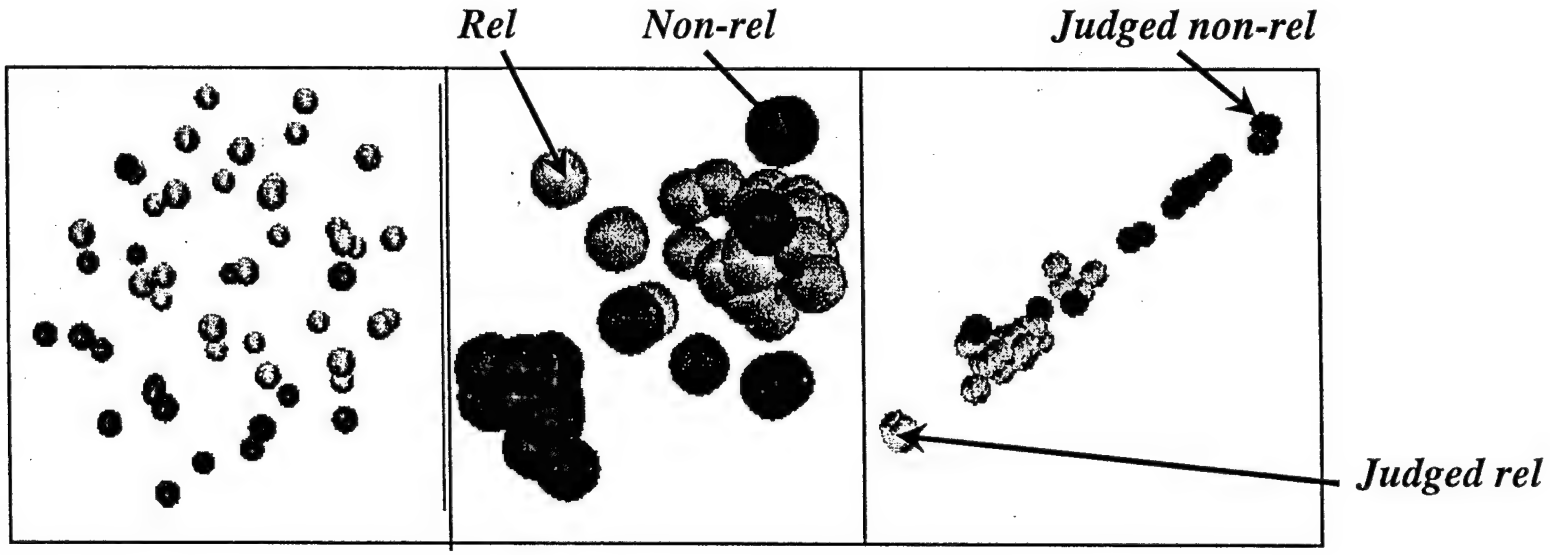
Browsing and Classification

- Developed 3-D graphics visualization tool for interactive browsing
  - First evaluation done in TREC this year
  - Aim is to demonstrate utility in improving search performance
  - Currently runs on SGI platform
- Downloaded PTO classification data
  - First version of testbed
  - Tasks defined
  - Demonstration system built

Aspect InQuery



Assisted Cluster Browsing



Browsing and Classification

- Refined categorization program for large databases
  - Previously tested with medical and essay data
  - Tested in recent TREC routing track
  - Nearest neighbor, Bayesian, Rocchio classifiers
  - Initial focus is on nearest neighbor

Image Indexing and Retrieval

- Goal: Develop similarity-based techniques for retrieving images such as trademarks, logos, and designs
  - Subgoal: Use both PTO and external data
- Technical Focus: Combine "appearance-based" approaches with simpler color and shape-based retrieval.
  - Subfocus: Develop multimodal techniques that can efficiently index and search very large databases of images

Image Indexing and Retrieval

- Developed new appearance-based image retrieval techniques
  - 50 times faster than previous for partial image matching, even faster for whole image matching
  - tested on general image data
- Downloaded PTO trademark and other miscellaneous images
  - converted Yellowbook and Trademark images to standard TIFF
  - Created subset (2000) of most recent trademarks for testbed
  - Created larger subset (>50,000) of non-text-based trademarks

Image Indexing and Retrieval

- Started creation of external "logo" database
  - higher quality, non-binary, color images
- Developed improved color retrieval technique
  - evaluated using color images from magazines such as logos and products
  - started evaluation of plant patent images
- Developed first version of shape-based retrieval
- Developed first version of image-based relevance feedback
- Developed multimodal demonstration systems

Distributed Retrieval Architecture

- Goal: Develop techniques for effectively and efficiently search very large, distributed databases
  - Subgoal: Use high-speed network as a demonstration platform
- Technical Focus: Extend and test current client-server architecture for multi-terabyte databases; Improve resource selection and result merging algorithms
  - Subfocus: Evaluate in TREC, with simulations, and on high-speed network

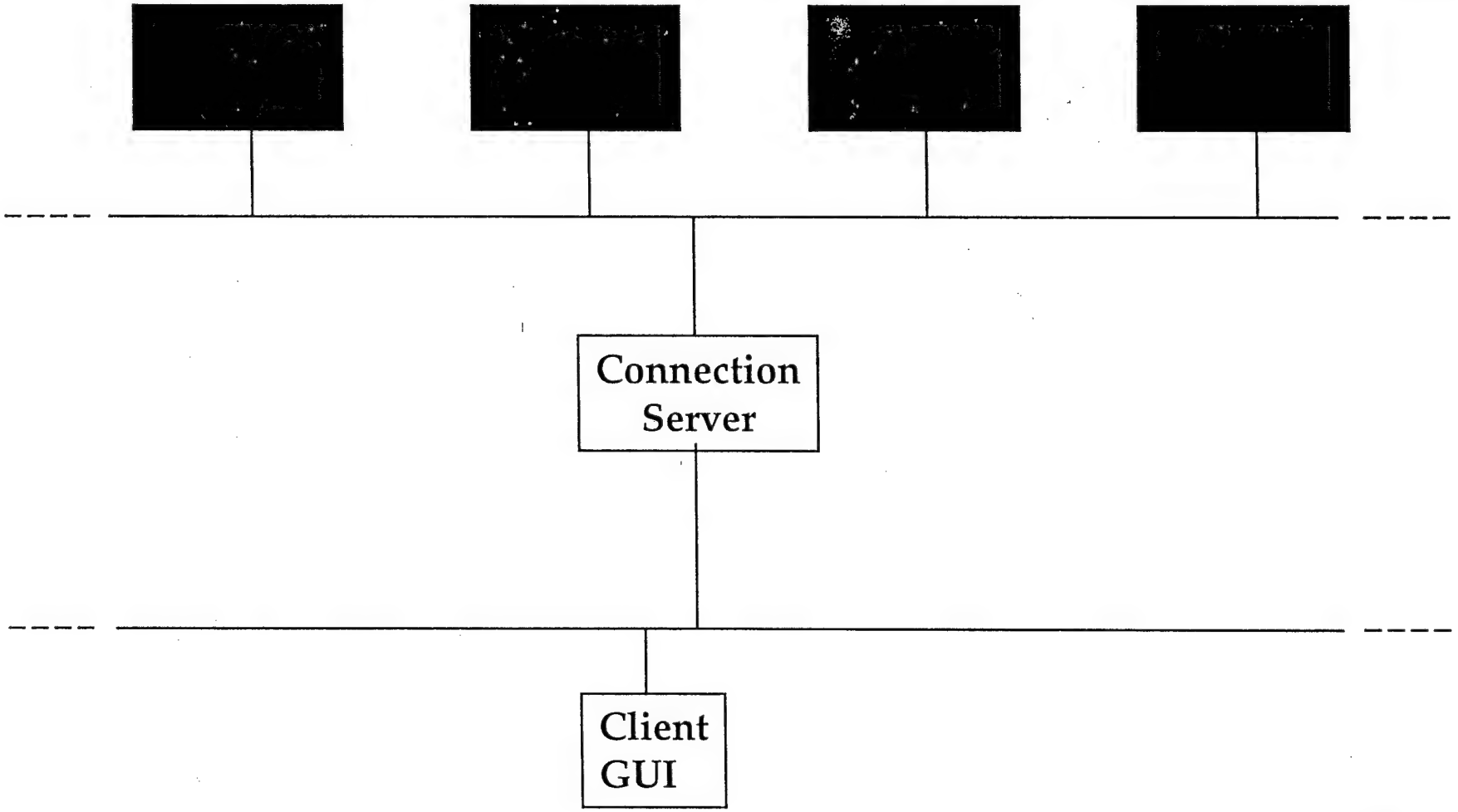
Distributed Retrieval Architecture

- Developed new approaches to resource selection and result merging
  - Results show resource selection is of primary importance
  - Simple word lists are effective as resource descriptions but may not scale
  - Other approaches being tested
- Studied indexing performance in large databases
  - evaluated in recent TREC very large corpus track

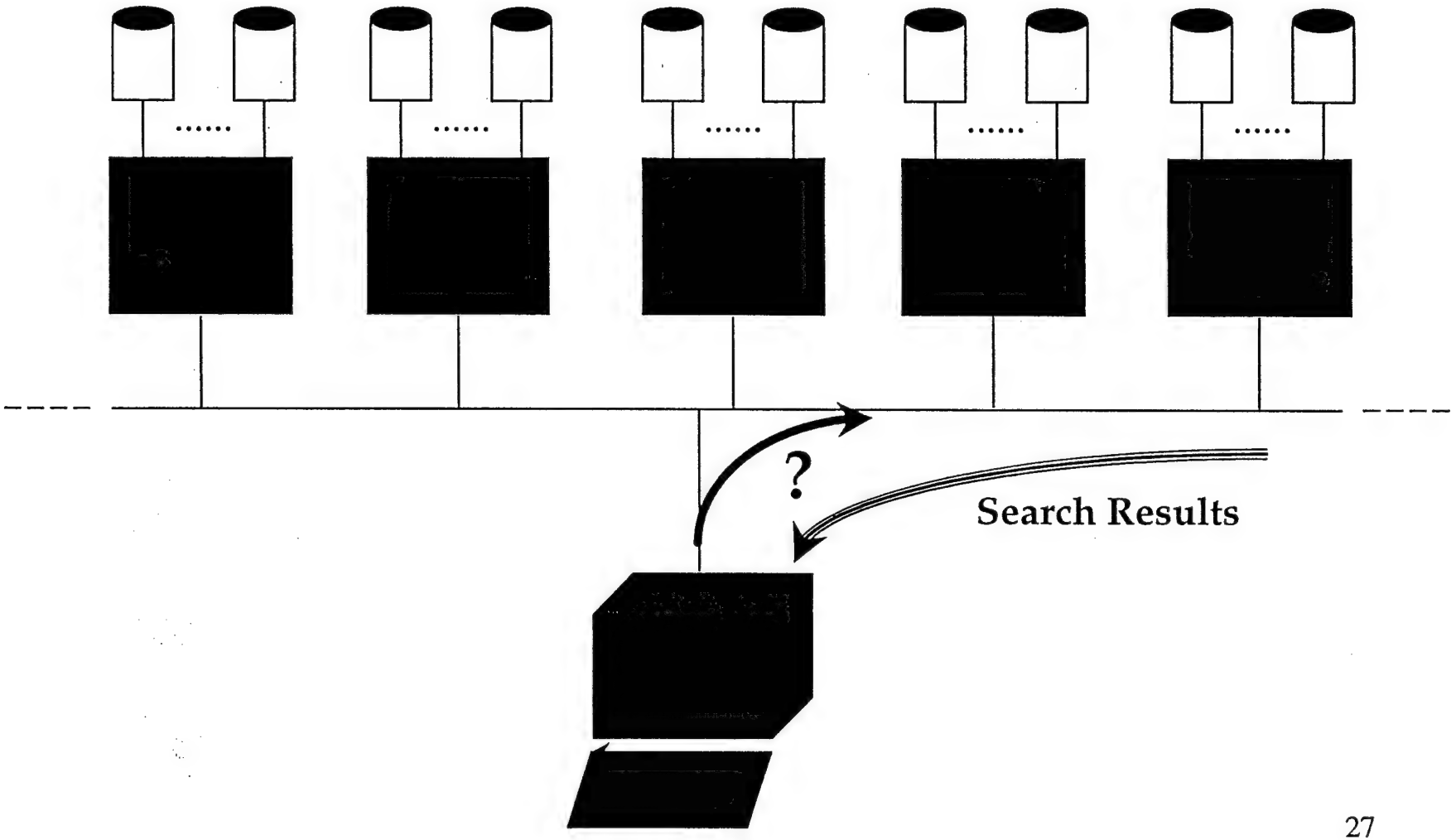
Distributed Retrieval Architecture

- Performed simulation studies of client-server architecture
  - tested three level architecture with clients, servers and "connection servers" currently implemented in INQUERY
  - moved functionality between layers to observe impact
  - used threaded and unthreaded implementations
- DS3 connection installed December 1997

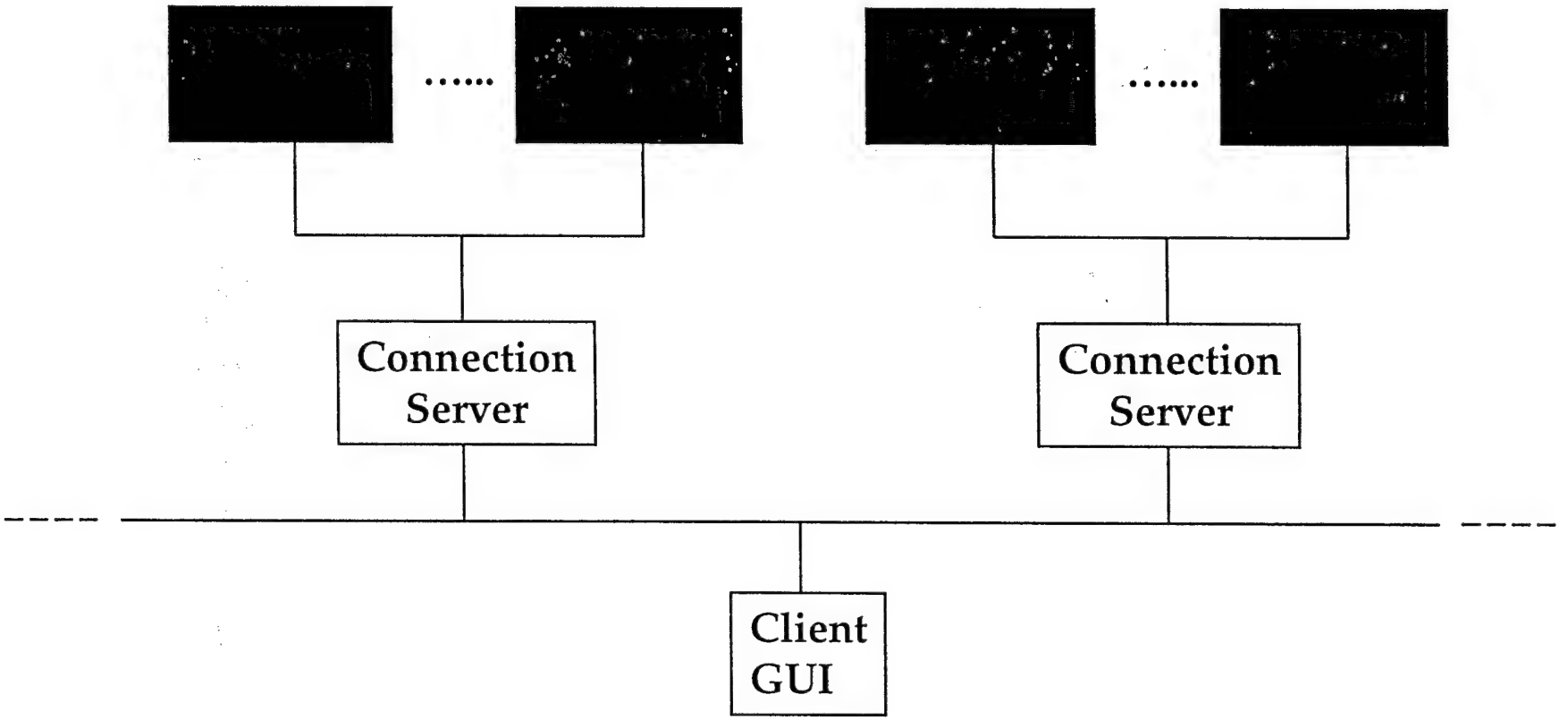
The INQUERY Distributed Architecture: Local Area Network



Distributed Heterogeneous Systems



The INQUERY Distributed Architecture: Wide Area Networks



Presentation Overview

- Patent Retrieval
  - presentation
  - demonstration
- Patent Classification
  - presentation
  - demonstration
- Patent Image Retrieval
  - presentation
  - examples of processing plant patent images
  - demonstration of multimodal trademark retrieval
  - demonstration of feedback and retrieval on other images

Patent Retrieval

Leah Larkey

Center for Intelligent Information Retrieval

University of Massachusetts, Amherst

December 10, 1997

Features

- All Patents from 1995 and 1996
  - 222,237 patents
- 50+ fields represented
- Queries
  - Unstructured

    I want technology that parents can use to control television content
  - InQuery operators

    #phrase(picture frames)

    #field(ASSG Microsoft)

Features

- Relevance Feedback
  - Retrieve docs based on user query
  - User marks a few good retrieved docs
  - System modifies query to get more docs like those marked
- Automatic Query processing
  - System adds phrases, compounds, related to query
- Suggest additional terms, phrases
  - System provides a list of possibly related terms
  - User may select some to add to query

Automatic Query Processing

- Add phrases using phrase dictionary built from database

  hot dogs → hot dogs #phrase (hot dogs)

- Add compounds using a general compound dictionary

  sun glasses → sun glasses #syn(#1 (sun glasses) sunglasses)

- Add compounds if hyphens

  in-line skates → in-line skates #phrase(in-line skates) #phrase (inline skates)

Overview of Patent Classification Projects

Leah S. Larkey

Center for Intelligent Information Retrieval
University of Massachusetts, Amherst
December 10, 1997

General Issues

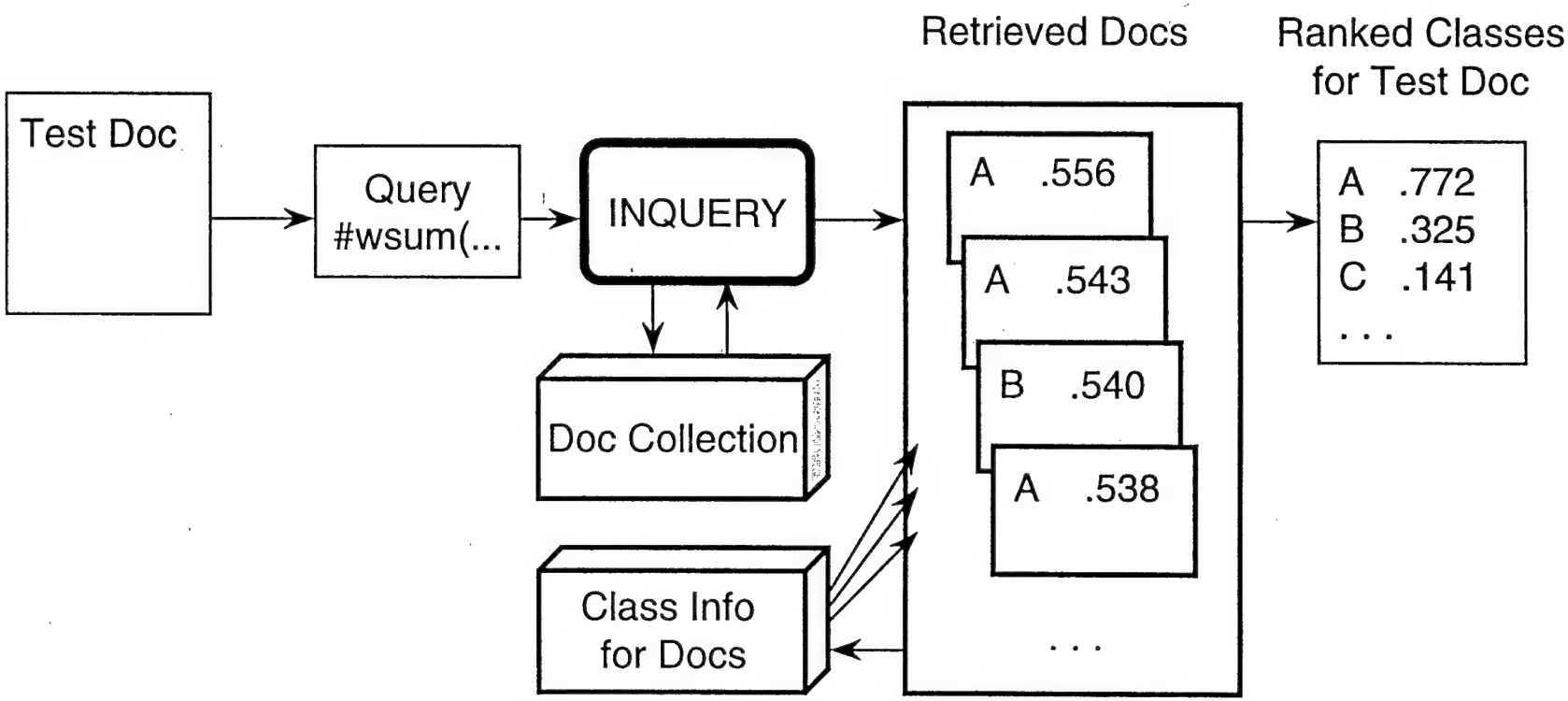
- Searching for Prior Art
  - Find relevant or similar patents to application
  - Find relevant non-patent literature
- Classification
  - Route patent application to correct Art Unit
  - Assign application to class and subclass
- Reclassification
  - Reorganization of existing class(es) into new subclass structure
  - Assignment of cross references after reclassification
  - Finding classes that need reorganization

Three Classifier Types

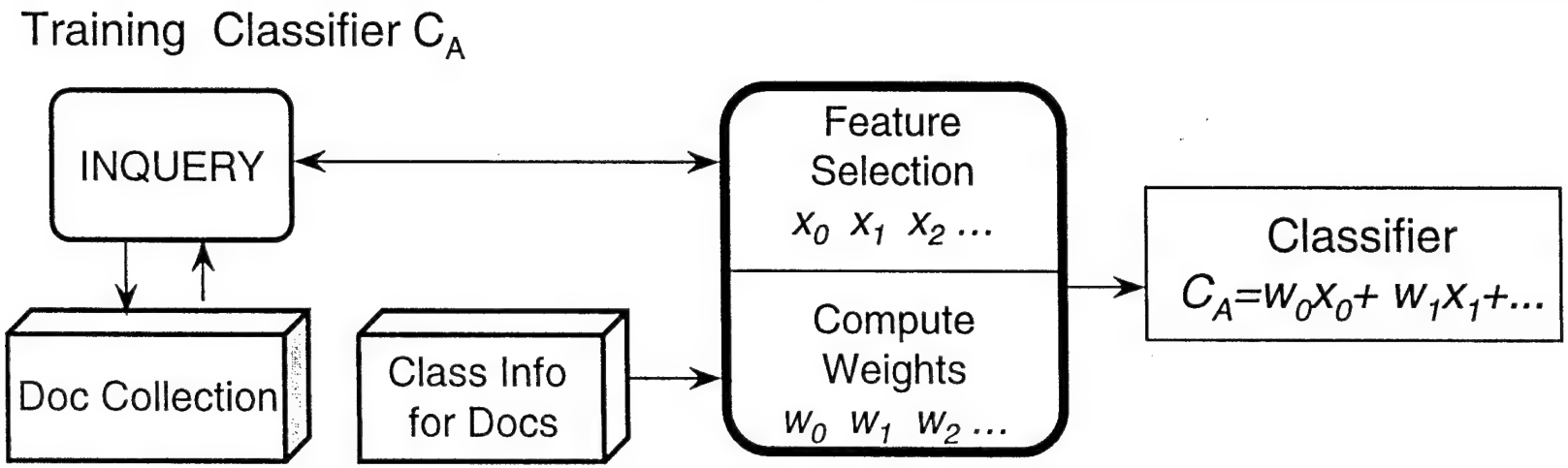
- *K*-Nearest Neighbor Classifier
- Bayesian Independence Classifier
- Relevance Feedback (Rocchio) Classifiers

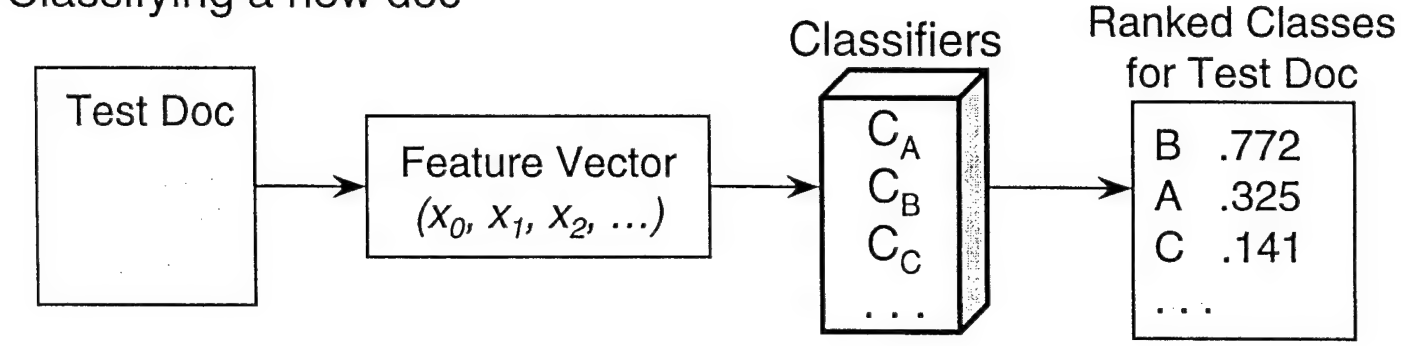
- Combinations of Classifiers

K-Nearest-Neighbor Classifier



Bayesian and Relevance Feedback Classifiers





Issues in *K*-Nearest Neighbor Classification

- Query Formulation
  - How to turn a document into a query.
- Deriving ranking scores for classes
  - Use scores and classes of retrieved documents to assign scores to candidate classes for test document.

Query Formulation

Query = weighted sum:

#wsum (1

w_{title} #sum ([Title])

$w_{abstract}$ #sum ([Abstract])

w_{bsum} #wsum ([most important Background Summary terms/phrases])

w_{detd} #wsum([most important Detail Description terms/phrases])

)

Query Formulation Example

Title: Adjustable skate brake
Abstract: The present invention comprises a brake having a slot formed in a support for receiving an adjusting screw which slidably secures the support to the skate. The head of the adjusting screw engages the surface of the ...
Detailed Description: Figs. 1-6 illustrate a first embodiment of an adjustable brake ...
Background Summary: In many present brake systems for both in-line skates and roller skates, a brake pad...

```
#wsum ( 1
  3.0 #sum(Adjustable skate brake)
  1.0 #sum(The present invention
  comprises a brake having a slot formed
  in a support for receiving an adjusting
  screw which slidably secures the
  support to the skate. The head of the
  adjusting screw engages the surface of
  the support ...)
  1.0 #wsum( 3 skate 3 brake 2 surface 2
  screw...)

  1.0 #wsum(5 adjustable 4 brake 2 skate
  1 bracket ...)
  )
```

Document Scores to Class Scores

$$class\_score_c = \sum_{i \in \text{ retrieved docs}} \left(doc\_score_i \cdot w_{i,c} \right) / n$$

$$w_{i,c} = \begin{cases} 0 & \text{if } c \text{ is not assigned to doc } i \\ 1 & \text{otherwise} \end{cases}$$

*K*NN Classification Example: Ranked List of Retrieved Documents

Query: #wsum(1 3 #sum(adjustable skate brake) 1 #sum([abstract]))

Retrieved Docs:

Patent	class/subclass	Title
5486011	280/11.2	Spring biased braking device for in-line roller skates
5487552	280/11.2	Braking mechanism for in-line skates
5505468	280/11.2	Braking device particularly for skates
5486012	280/11.2	Braking system for in-line skates
5549309	280/7.1	Multi-line in-line roller skate, ... roller skate frame
5524913	280/11.22	In-line pneumatic-tired roller skate with scrapers
5505469	280/11.2	Braking device particularly for skates
5482301	280/11.2	Self leveling in-line skate brake
5484149	280/11.26	Adjustable roller skate structure
5544026	362/103	Running lights for in-line roller skates

*K*NN Classification Example: Ranked List of Class/Subclass Candidates

Query: #wsum(1 3 #sum(adjustable skate brake) 1 #sum(<abstract>))

Ranked list of classes:

Subclass	Score	Description
280/11.2	.328	Land Vehicle/Skates: Wheeled Skates: With brake
280/7.1	.055	Land Vehicle/Convertible
280/11.22	.054	Land Vehicle/Skates: Wheeled Skates: Tandem Wheels
280/11.26	.054	Land Vehicle/Skates: Wheeled Skates: Extensible
362/103	.054	Illumination/ With wearing apparel or body support

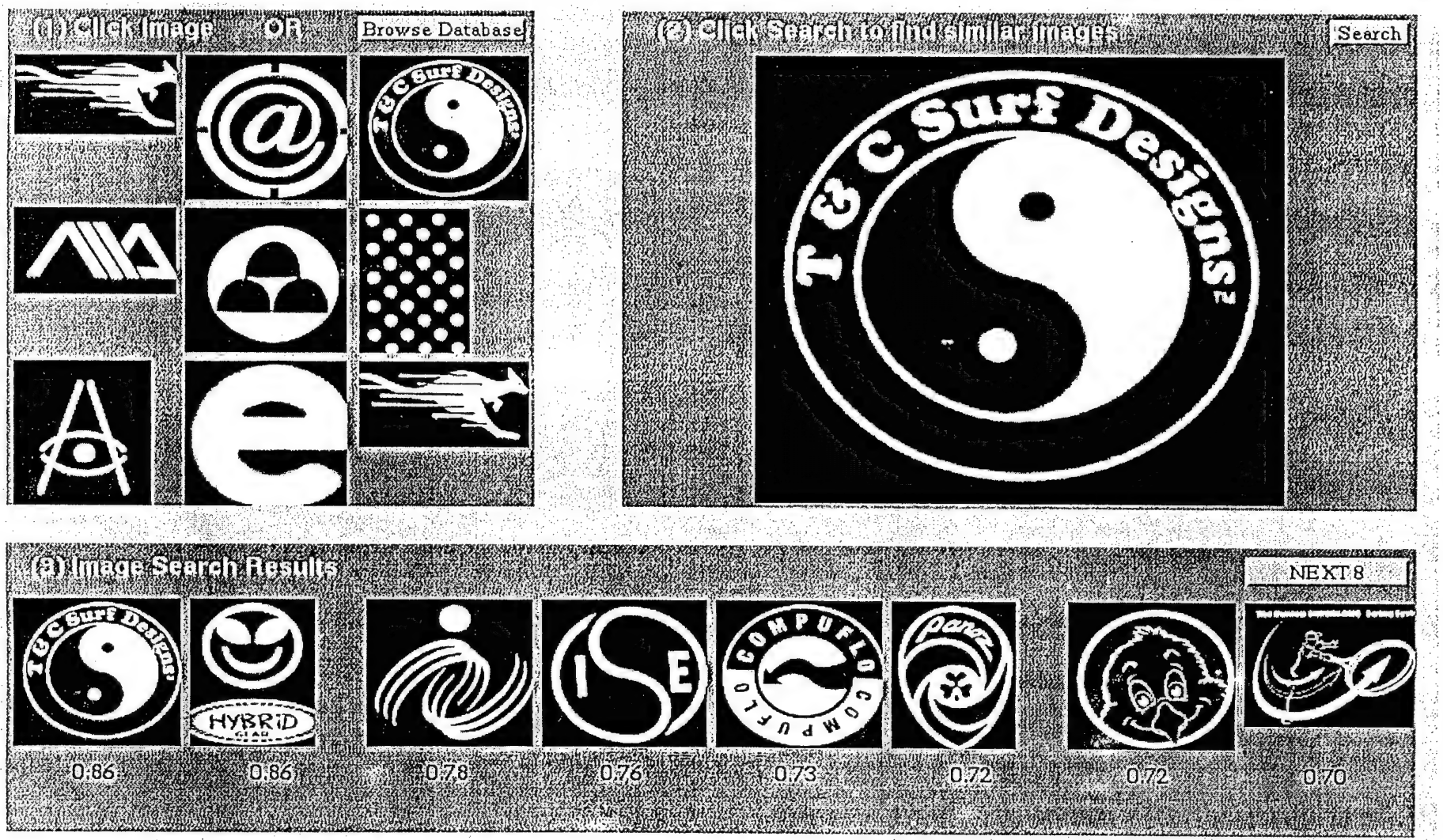
Image Retrieval

R. Manmatha
Center for Intelligent Information Retrieval
University of Massachusetts, Amherst
http://hobart.cs.umass.edu/~mmedia

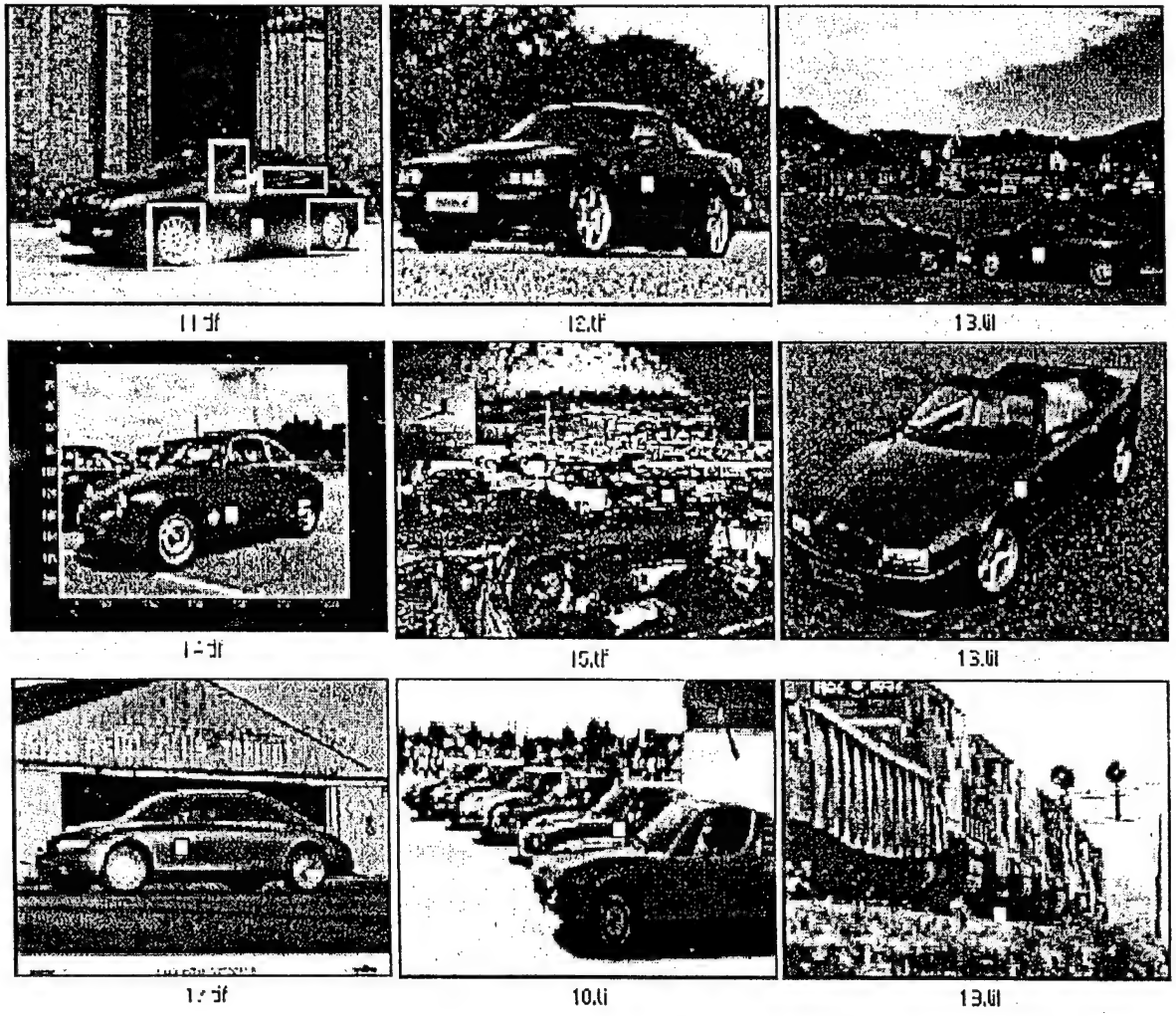
Image Retrieval

- People would like semantic answers to retrieval questions
  - Show me trademarks like this one
  - Are there designs that look like this?
  - Are there flowers with similar colors?
- Semantic retrieval hard to do.
- Retrieval based on similarity of image attributes
  - appearance, texture, color and shape .

Trademark Example



Example Pictures



Overview

- Appearance based image retrieval
  - Part image matching
  - Whole image matching.
- Relevance feedback
  - User input provided to improve results.
- Color based image retrieval.
- Combining Image and Text Retrieval for trademark retrieval.

Databases

- External database of 1561 greylevel images of cars, faces, apes etc. Some similarity to design patents.
  - obtained from the internet and cdroms.
- Trademark database of 63718 images from PTO
  - Images pre-processed by automatically cropping and reducing them.
  - image retrieval combined with text retrieval using INQUERY.
- Color database of advertisements which have product or brand logos
  - 400 images.

People Involved

- Chandu Ravela
- Thomas Michel
- Madirakshi Das
- Victor Wu
- R. Manmatha
- Edward Riseman

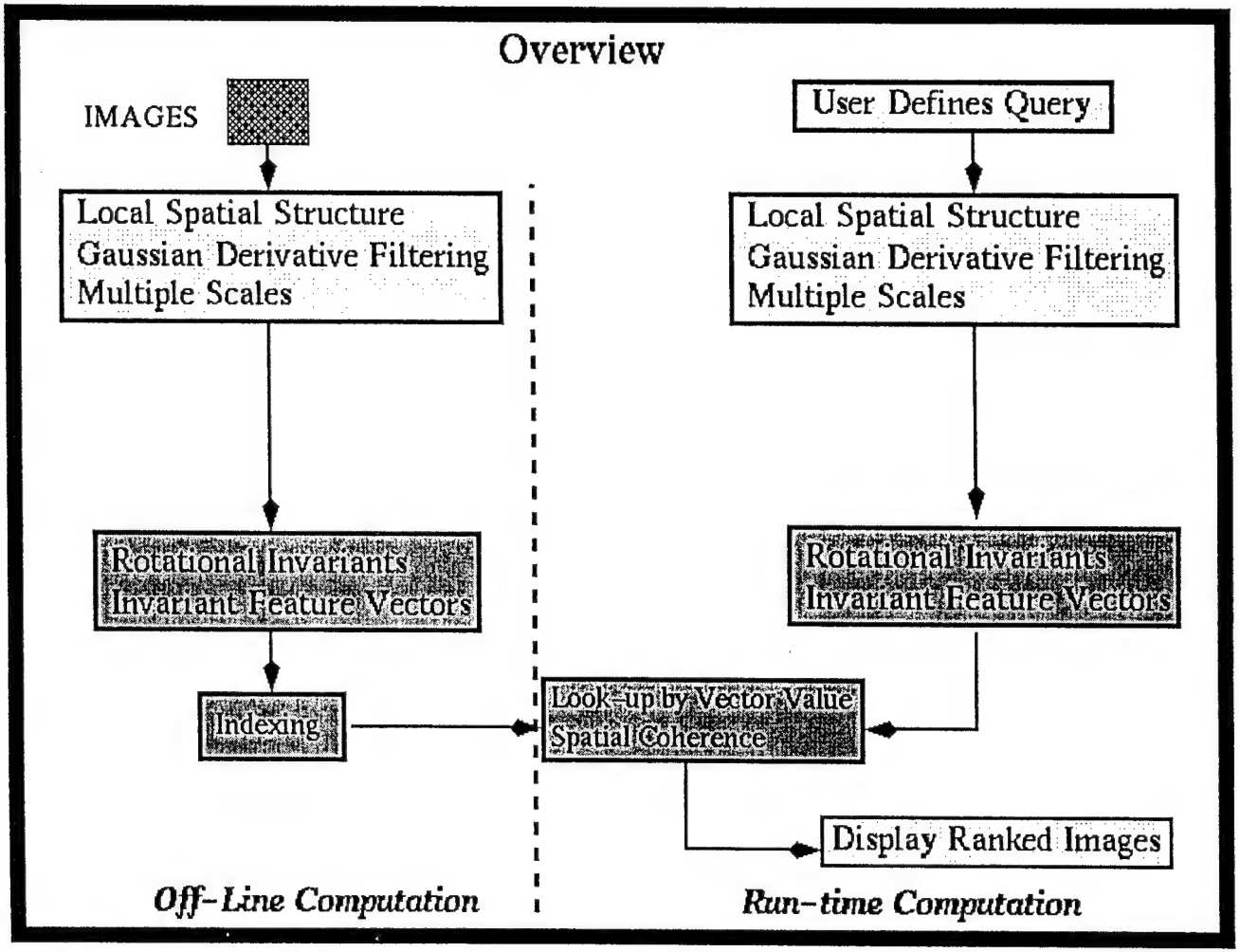
Publications

- Appearance based image retrieval.
  - 5 conf. papers. SPIE'97, SIGIR'97, CAIVL'97, DARPA IUW'97, ICCV'98.
  - 1 journal paper submitted to CVIU.
- Color based image retrieval.
  - 1 conf. paper in CVPR'97.

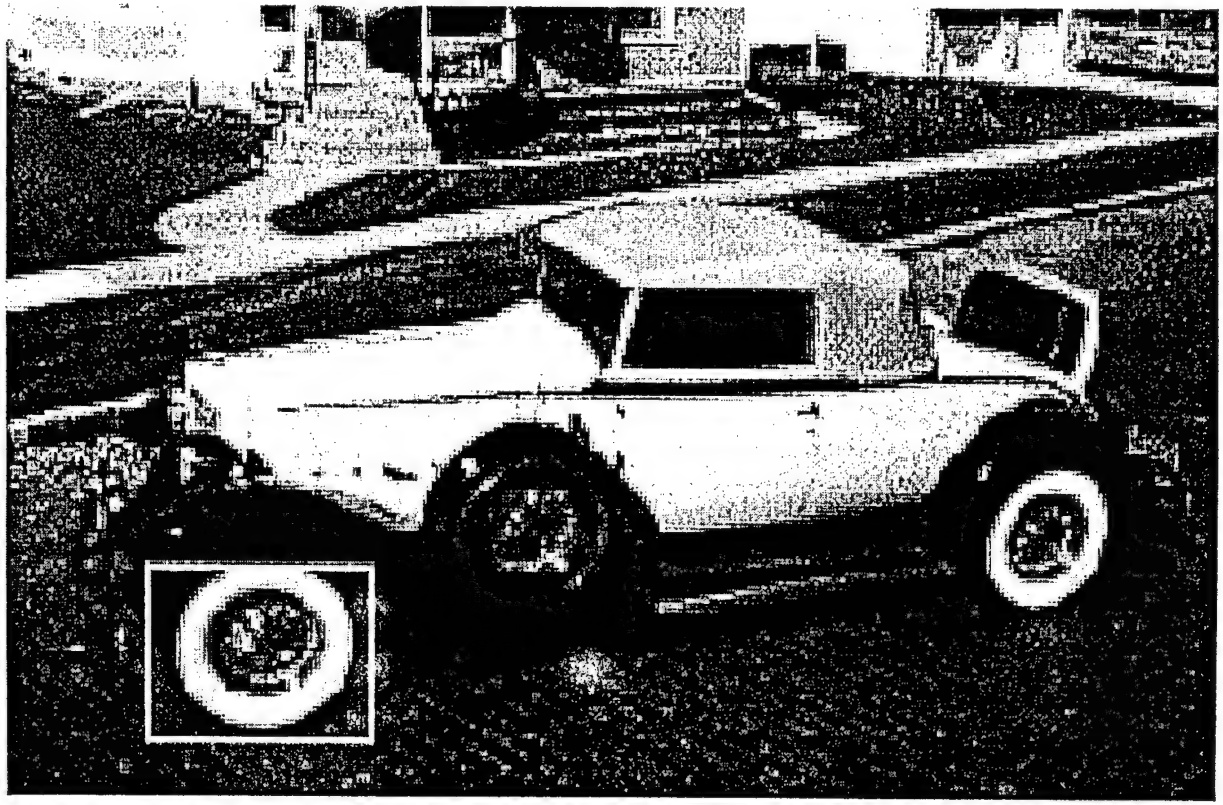
Part Image Retrieval

- User outlines query
- Query is matched to database images.
- Database images ranked according to similarity.
- Advantages:
  - Image may be embedded against arbitrary backgrounds.
  - View variations up to 25 degrees tolerated.
  - No learning required.
- Disadvantages: Slow
  - speeded up 50 times but still takes from 1 to 7 min.

Part Image Matching



Car Query



Results of Car Query

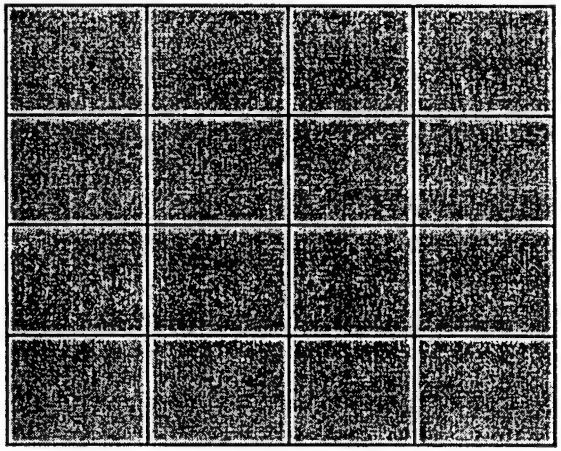
Whole Image Retrieval

- Find and rank images in the database which are similar to the example image.
- Advantages: Fast.
- Disadvantages: Not able to handle parts of images.
- May be based on different features:
  - Moments.
  - Jpeg coefficients.
  - Curvature, phase.

Moments

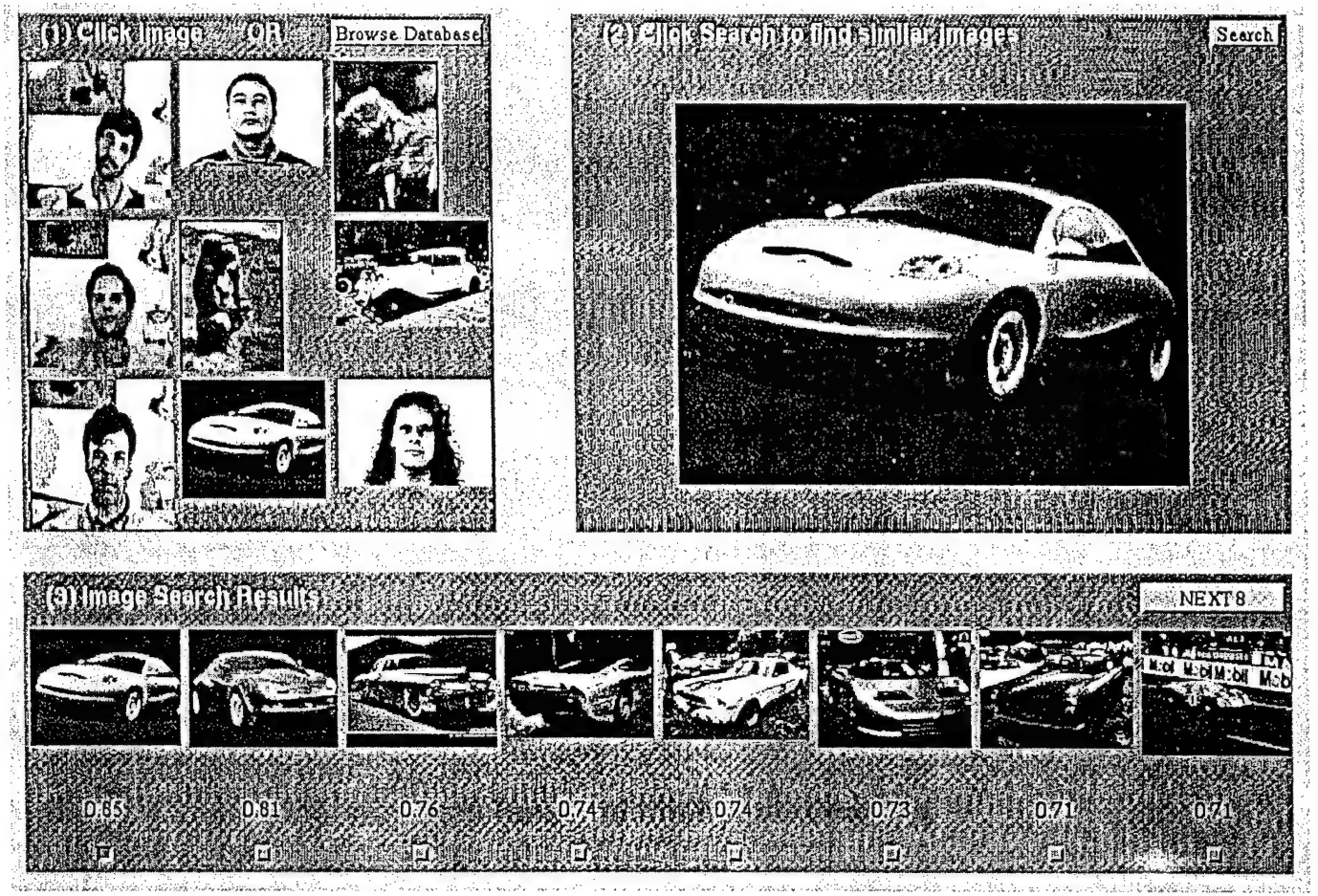
- Traditionally used to characterize shape.
- Our experience - poor features to use.
- Results poor.

Jpeg Matching.

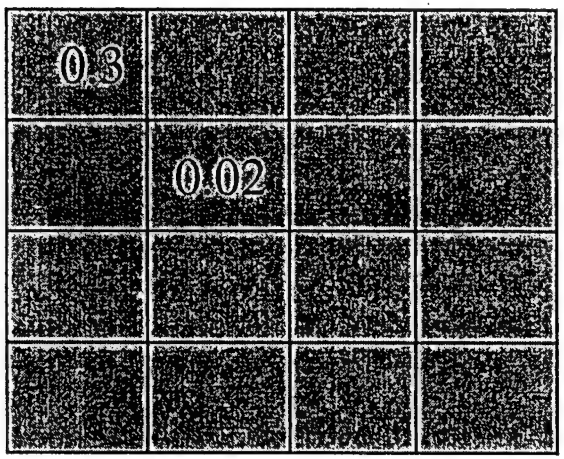


- Jpeg image divided into 8 by 8 blocks.
- Jpeg coefficients available for each block.
- Match images by comparing jpeg coefficients for corresponding blocks.
  - Sum over all blocks and use as error measure.

Jpeg Retrieval

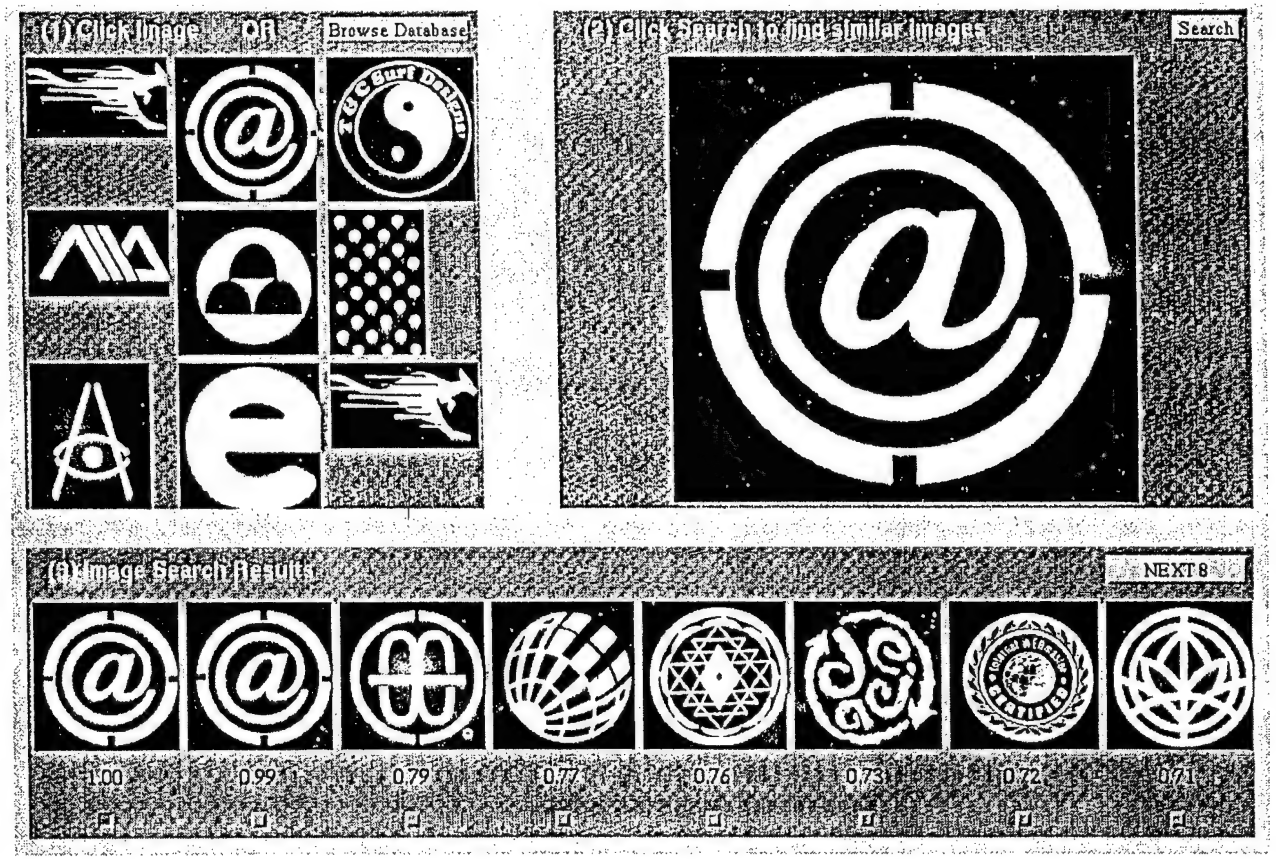


Relevance Feedback

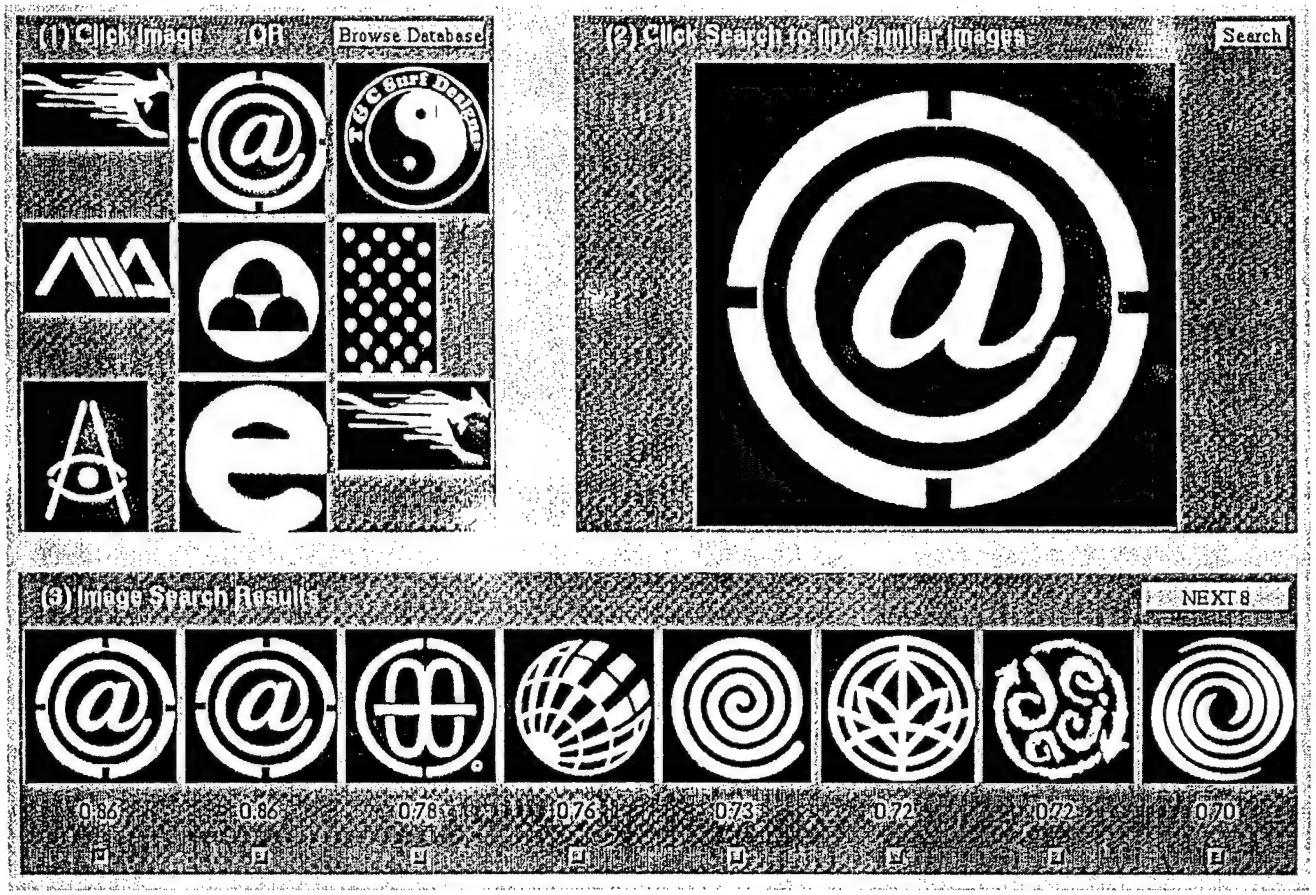


- User feedback used to modify similarity search.
- From the retrieved images user specifies which ones are relevant.
- Blocks weighted differently when computing error measure.
- Blocks common to relevant images are weighted more than blocks common to non-relevant images.
- Different weights for different coefficients

Jpeg Retrieval

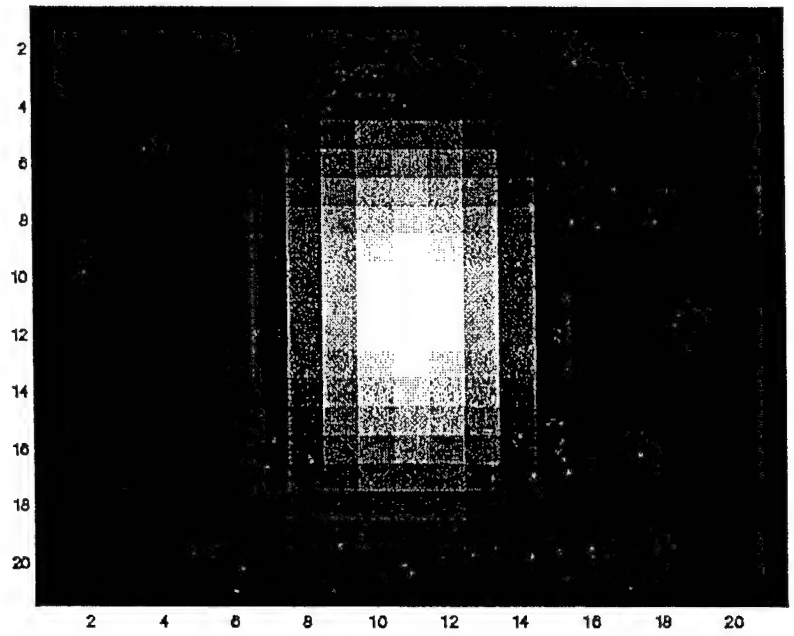


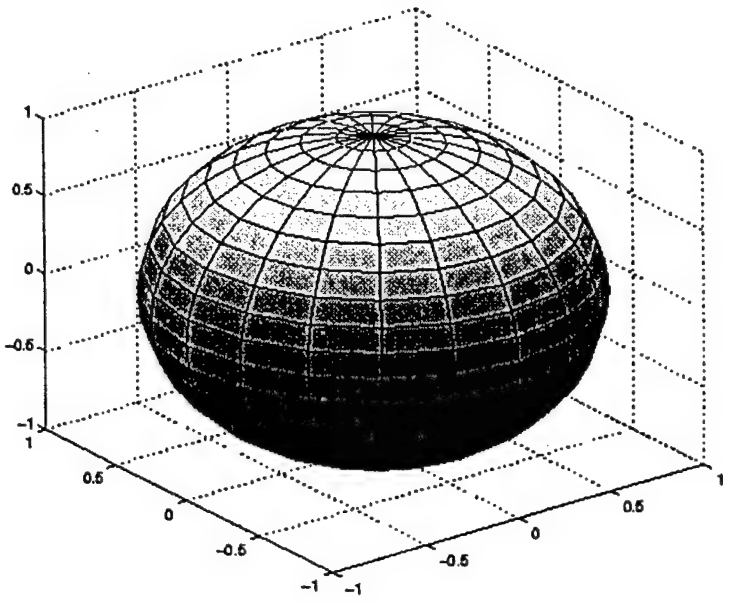
Jpeg Retrieval



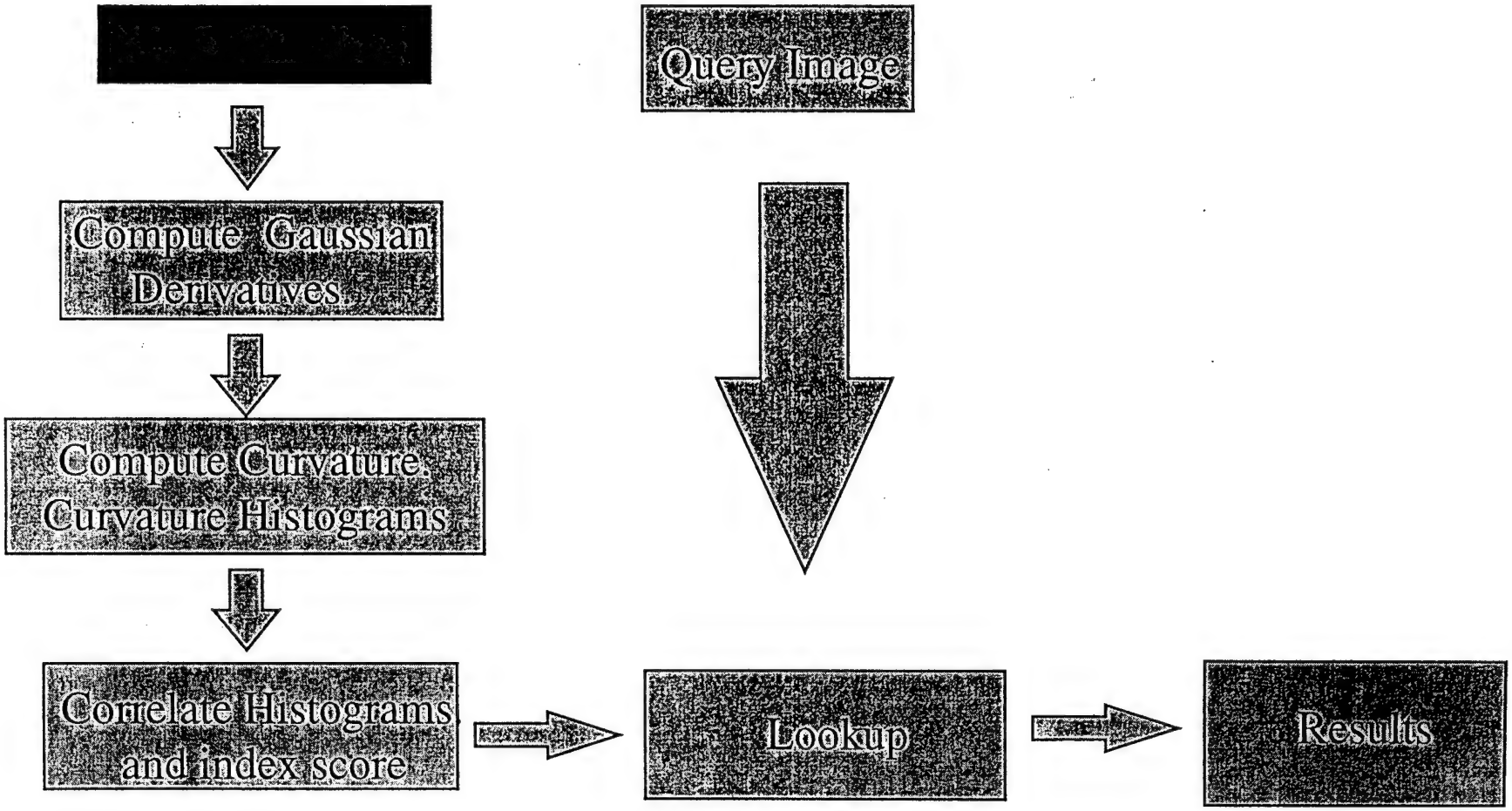
Local Curvature

- Local curvature - a good description of the surface locally.





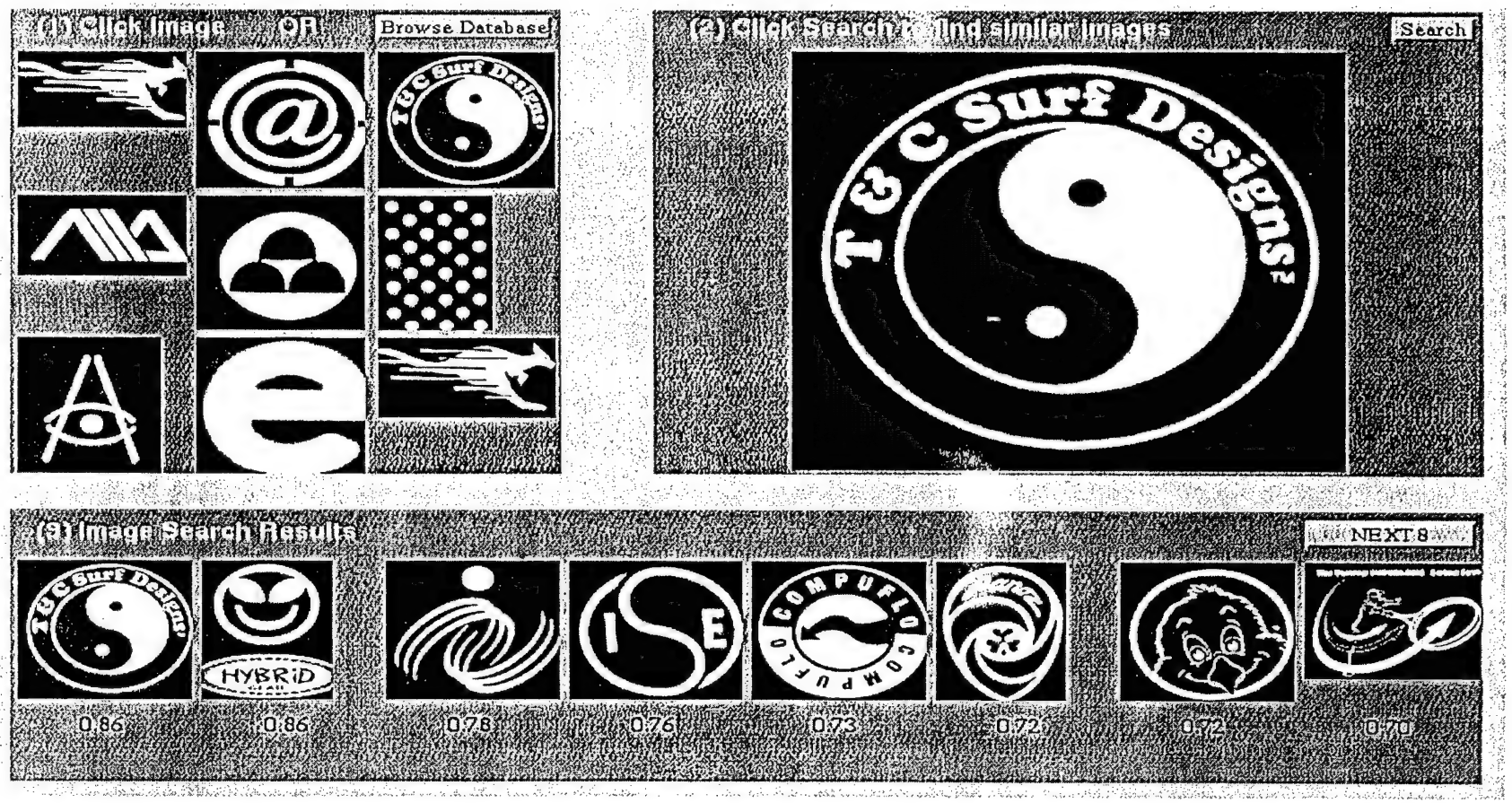
Curvature Matching



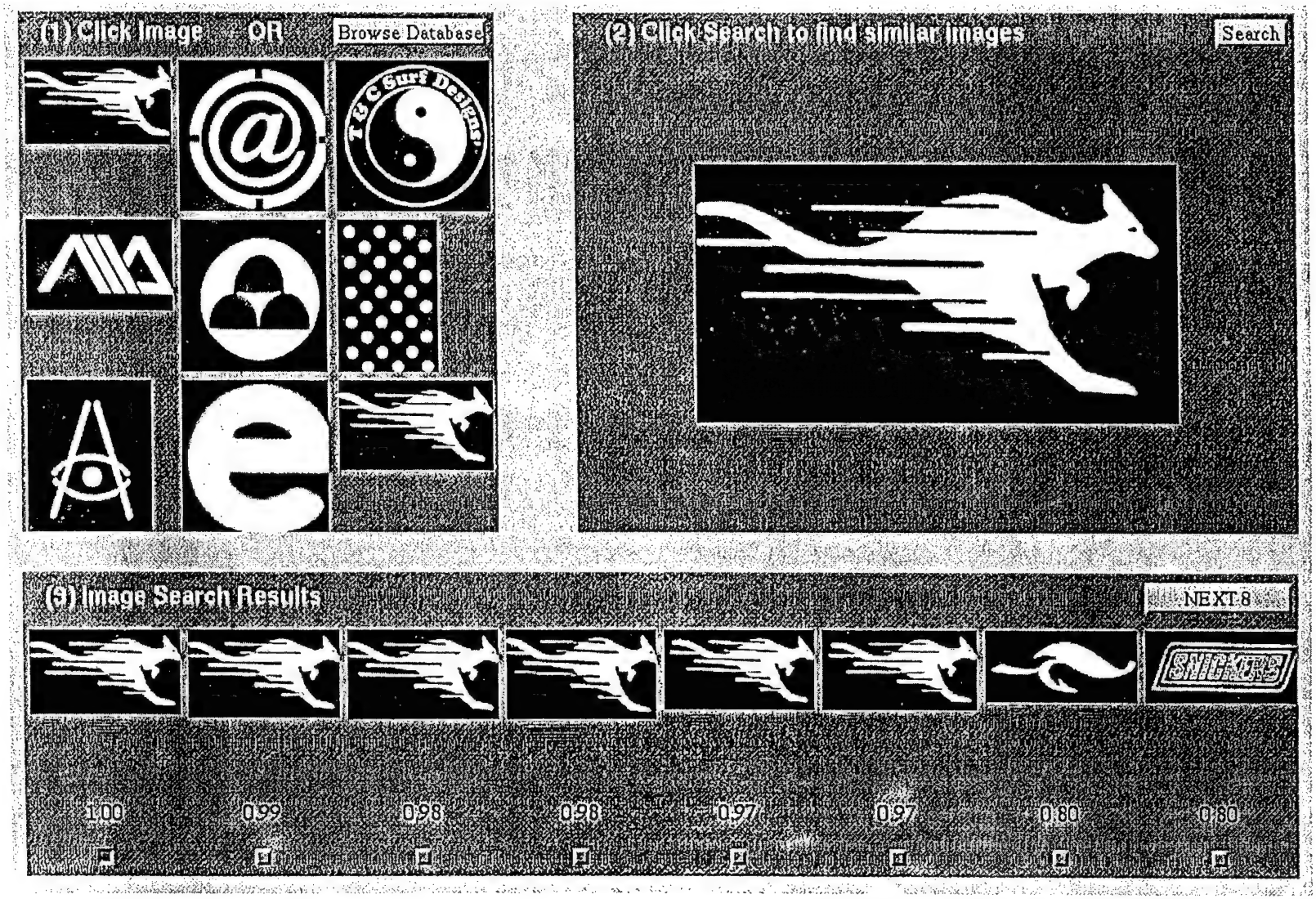
Phase Matching

- Phase angle may also be used as a feature.
- Roughly - signature of how many edges at what orientations.
- Use phase histograms as for curvature.
- Use of features may be database dependent.
- Combine curvature and phase.

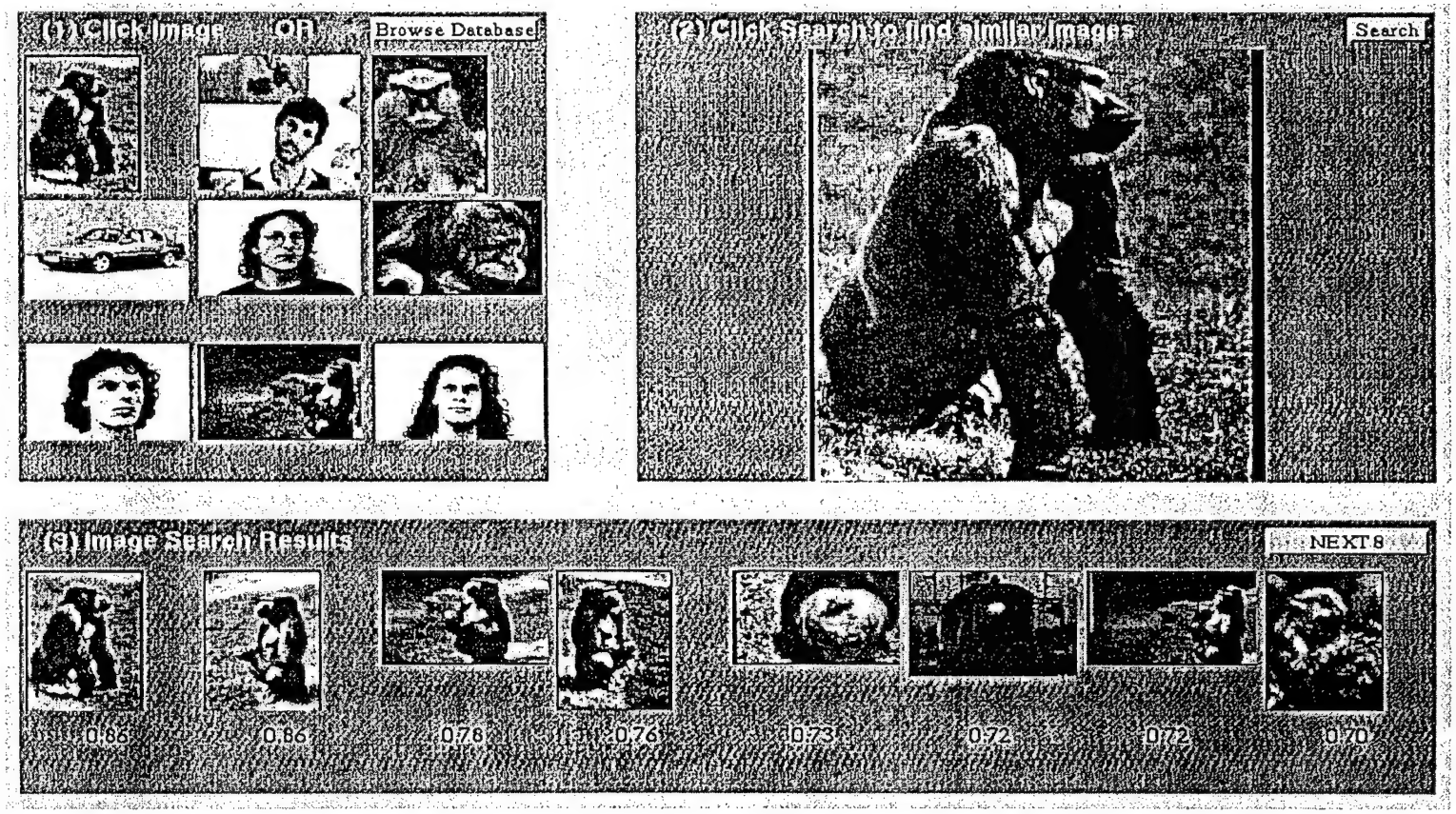
Curvature - Results.



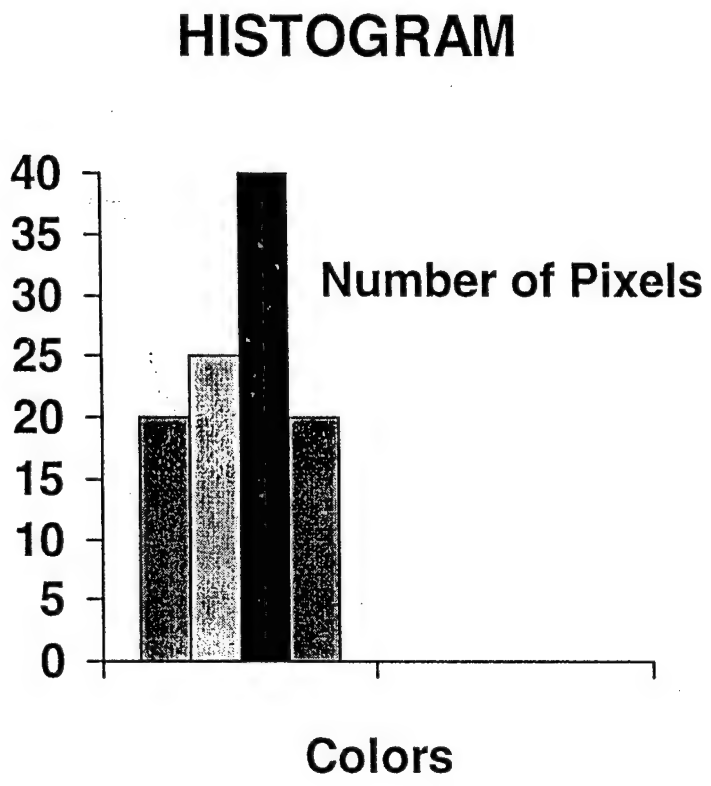
Curvature - Results



Curvature - Results.

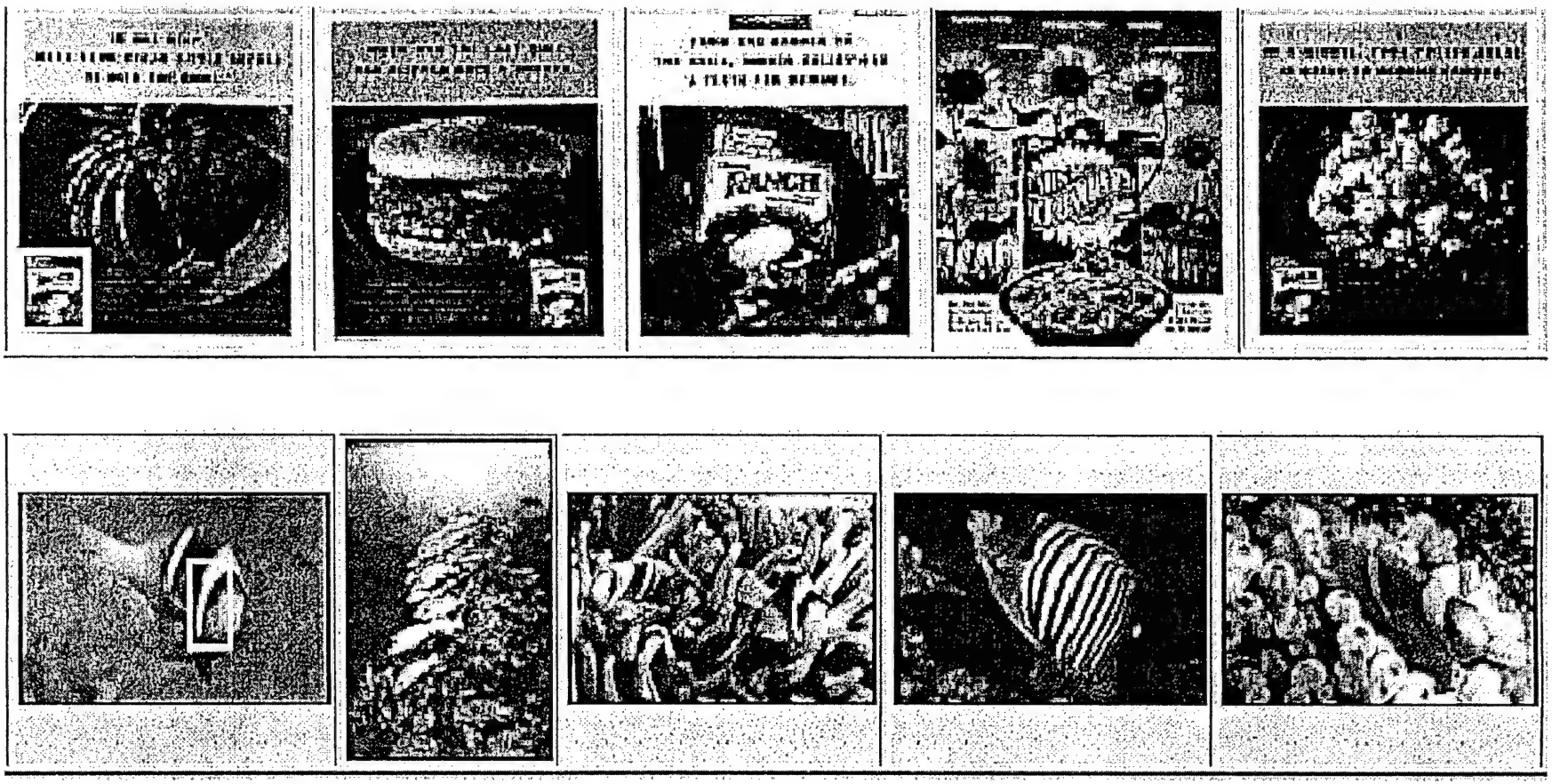


Color Image Retrieval



- Retrieve images similar in color.
- Compute local color histograms.
- Compute spatial adjacency graph
  - specifies which colors are adjacent.
- Specify query using mouse.
- Database
  - 400 images of advertisements. Search done using company or brand logos.
  - 800 general images from cdrom.

Results of Color Retrieval



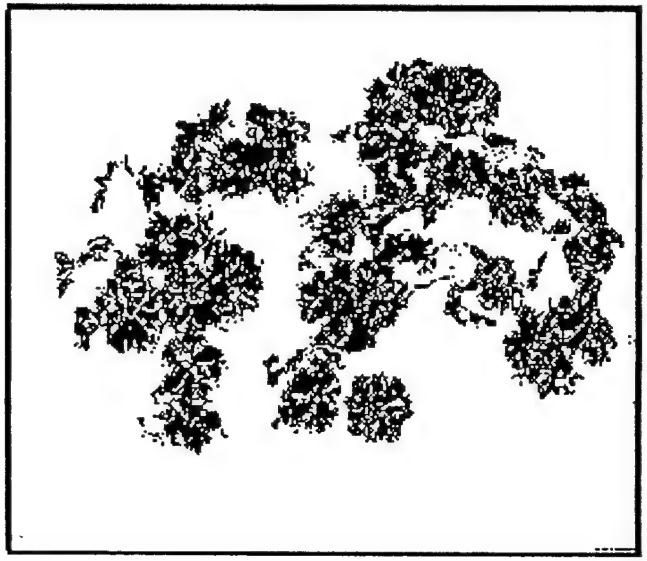
Plant Patents

- Possible Approach
- Color of flower
  - Color histogram of flower region.
  - Color adjacencies in this region.
- Text Information
  - Plant description.
  - Flowering description.
  - Propagation Methods.

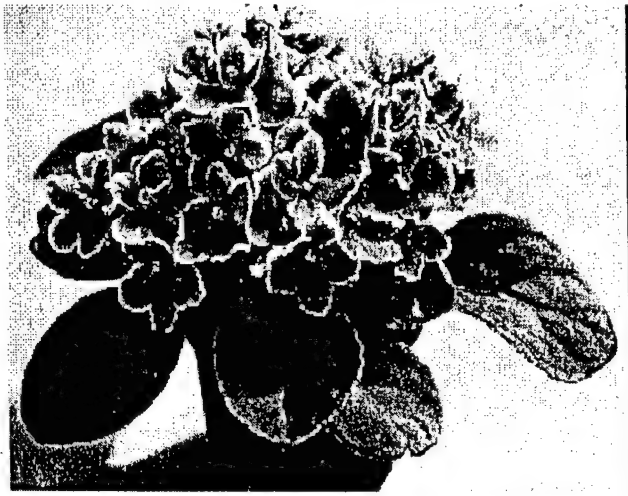
Plant Patents

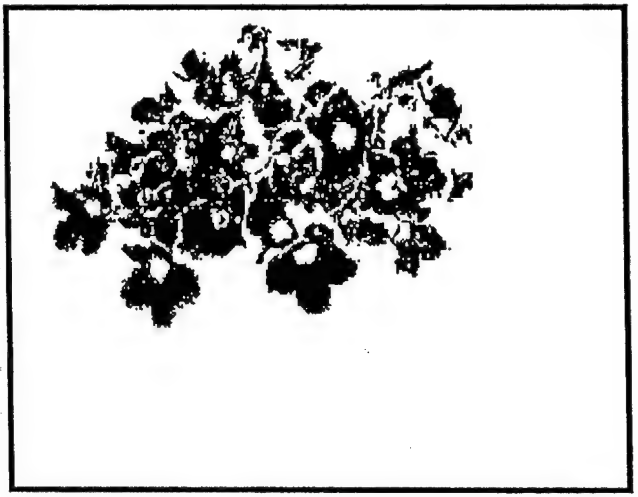
Extraction of 'flower' regions

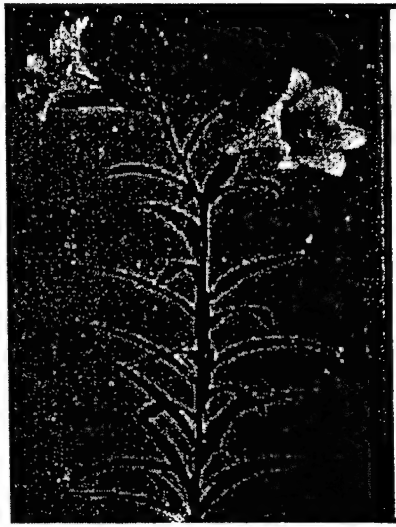
- Eliminating background
- Eliminating leaves
- Finding regions of significant size in remaining image



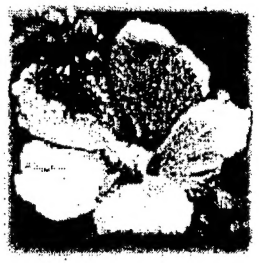
Plant Patents







Example Web Page



- Kyewolhyang
  - Single and bell-shaped
  - Light purple flowers with small red eye.
  - Branches upright.
- Koyoro
  - Light pinkish purple flower with small red eye.
  - Very short radiate vein. Broad and round petals glabrous.
  - Mid-season blooming type. Branches upright.
  - The meaning is 'calm'.
- http://www.ssc.samsung.co.kr/ss/

Image and Text Retrieval

- Retrieval based on image content may not be able to retrieve certain items.
  - Who took this photograph?
  - Stylized pictures or pictures with radically different viewpoints.
  - Textual information (eg Coca Cola).
- Textual annotations may provide some of this information.

Image and Text Retrieval

- Initial retrieval using trademark classifications.
- Use on the the images for Image retrieval using curvature and phase
  - an indexable demo - scores computed on the fly.

Future Work

- Improve indexing - strategies for scaling.
- Combining image and text retrieval scores.
- Search of web images:
  - comparison of web images with stored database.
  - need high speed data connection.
- Other features for appearance and shape.
- Use of relevance feedback over multiple methods ie. weight each method differently according to user feedback.
- Improve speed of part image techniques.